Genetic Variation Within and Among Populations of *Arabidopsis thaliana*

Joy Bergelson,* Eli Stahl,* Scott Dudek† and Martin Kreitman*

**Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637 and †Department of Biology, Washington University, St. Louis, Missouri 63130*

Manuscript received July 1, 1997
Accepted for publication November 17, 1997

ABSTRACT

We investigated levels of nucleotide polymorphism within and among populations of the highly self-fertilizing Brassicaceous species, *Arabidopsis thaliana*. Four-cutter RFLP data were collected at one mitochondrial and three nuclear loci from 115 isolines representing 11 worldwide population collections, as well as from seven commonly used ecotypes. The collections include multiple populations from North America and Eurasia, as well as two pairs of collections from locally proximate sites, and thus allow a hierarchical geographic analysis of polymorphism. We found no variation at the mitochondrial locus *Nad5* and very low levels of intrapopulation nucleotide diversity at *Adh*, *Dhs1*, and *Gpa1*. Interpopulation nucleotide diversity was also consistently low among the loci, averaging 0.0014. g_{st}, a measure of population differentiation, was estimated to be 0.643. Interestingly, we found no association between geographical distance between populations and genetic distance. Most haplotypes have a worldwide distribution, suggesting a recent expansion of the species or long-distance gene flow. The low level of polymorphism found in this study is consistent with theoretical models of neutral mutations and background selection in highly self-fertilizing species.

THE estimation of nucleotide diversity in highly self-fertilizing species (Hamrick and Godt 1989) is of considerable theoretical interest to population geneticists. Self-fertilization, a form of inbreeding, is expected to reduce within-population diversity by a factor $(2 - s)/2$, where s is the selfing rate (Pollack 1987). With complete selfing, therefore, polymorphism will be reduced by a factor of two compared to an equivalent outcrosser. Since recombination is effectively reduced by self-fertilization (Golding and Strobeck 1980), a further reduction in neutral polymorphism is expected from genetic hitchhiking accompanying selective sweeps (Maynard Smith and Haigh 1974) and from background selection against deleterious mutations (Charlesworth *et al.* 1993). Background selection alone may be expected to reduce neutral variability by a factor of 10 in a selfer (Charlesworth *et al.* 1993). Reductions of silent polymorphism levels of this magnitude have been observed in regions of reduced recombination in Drosophila (Berry *et al.* 1991; Langley *et al.* 1993; Wayne and Kreitman 1996).

The reduction in nucleotide diversity is not expected to be equally apportioned within and among populations. Recent theoretical work indicates that local demic selection, as well as background selection, enhances interpopulational differentiation at neutral linked sites, and this effect is stronger in selfing populations (Charlesworth *et al.* 1997; Nordborg 1997). Self-fertilization may also increase interpopulation differentiation by reducing pollen dispersal, one of the two forms of gene flow among plant populations. Theory also suggests a higher signal-to-noise ratio (elevated neutral polymorphism compared to background level) and hence greater detectability of balancing selection under reduced recombination. Neutral polymorphism linked to selected alleles is expected to accumulate over a larger region than under higher recombination (Hudson and Kaplan 1988). Simulations incorporating the effects of background and balancing selection under partial self-fertilization confirm the expectation of a higher signal-to-noise ratio for a selectively maintained polymorphism (Nordborg *et al.* 1996). Nordborg *et al.* (1996) predict that in a selfer, DNA sequence polymorphism at a locus under balancing selection will far exceed that at a selectively neutral locus. Furthermore, this variation will be distributed largely between balanced haplotypic classes.

Estimated outcrossing rates of 1% or less (Redei 1975; Abbott and Gomes 1989) make *Arabidopsis thaliana* well suited for studying the impact of reduced effective recombination on the levels and patterning of nucleotide variation across the genome. At the present time, the greatest amount of information about nucleotide diversity in this species comes from efforts to establish genetic maps using naturally occurring polymorphism in accessions derived from natural populations (Reiter *et al.* 1992; Konieczny and Ausubel 1993; Vos *et al.* 1995). The effort to build molecular genetic maps in *A. thaliana* has been impeded, however, by an apparent lack of

Corresponding author: Joy Bergelson, Department of Ecology and Evolution, University of Chicago, 1101 E. 57th St., Chicago, IL 60637. E-mail: jbergels@midway.uchicago.edu

TABLE 1

Populations surveyed in this study

Name	Geographic origin	No. of lines	Year collected
BG	Seattle, WA	10	1993
Dem	Demotte, NJ	10	1993
NC6	Durham County, NC	10	1992
NC7	Durham County, NC	11	1992
RP	Ithaca, NY	11	1993
Got	Göttingen, Germany	10	1993
Kz	Karagandy, Kazakhstan	11	1994
NFC	Ascot, England	10	1993
NFE	Ascot, England	12	1993
Pu2	Prudka, Czechoslovakia	10	1991
Tamm	Tammisaari, Finland	10	1989

nucleotide polymorphism. For example, Konieczny and Ausubel (1993) had to use as many as 83 restriction enzymes to find a single polymorphic marker in 1.5–2-kb stretches of Columbia (Col) and Landsberg *erecta* (L*er*) strains, and most of the polymorphic markers they did find were at low frequency among other ecotypes (Hanfstingl *et al.* 1994).

Curiously, this lack of nucleotide polymorphism is not apparent in published surveys of genetic variation among *A. thaliana* ecotypes. Three studies have used RFLP analysis to estimate nucleotide diversity in worldwide collections of ecotypes, and all report high levels of nucleotide diversity that exceed 0.01 per nucleotide site. As we explain later, all these studies overestimate the true value of nucleotide polymorphism (see discussion). In addition, two studies have used sequence data to estimate nucleotide diversity. First, sequence polymorphism among 18 ecotypes was studied for two nuclear cleared-amplified polymorphic-sequences loci (Hardtke *et al.* 1996), and nucleotide diversity for the two loci averaged 0.0221. This is again a relatively high estimate although it is inflated by the inclusion of length changes. Second, nucleotide diversity of *Adh* and a 5′ flanking region was estimated from 17 ecotypes (Inan *et al.* 1996), and overall nucleotide diversity was estimated to be 0.0080. Since *Adh* is thought to be under balancing selection (Hanfstingl *et al.* 1994), it is unclear whether this estimate is representative of polymorphism levels. At the present time, there are few, if any, unbiased estimates of nucleotide polymorphism levels among ecotypes of *A. thaliana*. Furthermore, because the above-mentioned studies do not include multiple representatives from given sites, we know virtually nothing about the relative levels of nucleotide diversity within and among populations on a worldwide scale.

In this paper, we provide estimates of nucleotide polymorphism at one mitochondrial and three nuclear loci among 11 worldwide populations of *A. thaliana* using a PCR four-cutter RFLP technique. Multiple lines were field collected from local populations, allowing intra- and interpopulation comparison. The collections include multiple populations from North America and Eurasia, as well as two pairs of collections from locally proximate sites, and thus allow a hierarchical geographic analysis of polymorphism. In addition, we estimate levels of polymorphism among seven commonly used ecotypes and compare that estimate to naturally occurring variation. We find no variation at the mitochondrial locus and consistently low levels of polymorphism at the three nuclear loci; the largest fraction of variation segregates between populations. Most haplotypes have worldwide distributions, indicating gene flow over long distances, and we find no evidence of isolation by distance.

MATERIALS AND METHODS

Sample collections: Mature siliques were field collected from 10–12 randomly selected individuals in each of 11 natural populations to establish our sample. A total of 115 lines were established. The geographic locations of the collections are listed in Table 1. Five of these populations are located in North America (BG, Dem, NC6, NC7, and RP), and six are located in Eurasia (Got, Kz, NFC, NFE, Pu2, and Tamm). Two pairs of sites (NC6–NC7 and NFE–NFC) were selected to be close neighbors, separated by <1 km. This sampling strategy allowed us to explore geographic patterns of nucleotide polymorphism, both within and between populations, for populations collected in a hierarchical design. Our sample collections, in contrast to ecotypes from the Arabidopsis Biological Resource Center, Ohio State University (Columbus, OH) could not have been subjected to artificial selection. We considered an additional seven commonly used ecotypes (Col-0, L*er*, Tsu-0, Wu-0, No-0, Nd-0, Ms-0) that allow comparison of ecotypes to between-population samples and comparison of our study to other published studies. All our field-collected lines have been donated to the Arabidopsis Biological Resource Center.

Surveying nucleotide polymorphisms: DNA was extracted from one offspring originating from each maternal line according to a modification of published procedures (Cocciolone and Cone 1993; Colosi and Schaal 1993). Two or three leaves from each rosette were frozen in liquid nitrogen and ground with ball bearings. Frozen leaf tissue was thawed in 0.6 ml of urea lysis buffer and agitated at 37° for 10–60 min. After adding 0.5 ml of phenol:chloroform (1:1), the solution was agitated for 10 min and spun at room temperature for 5 min. Thirty microliters of 3 M NaOAc (pH 5) and 600 μl of isopropanol were added to the aqueous portion, and the solution was spun at room temperature for 3 min. The resultant pellet was then washed with 70% ethanol, air dried, and suspended in 50 μl TE (10 mm Tris, 0.1 mm EDTA, pH 8.0). The DNA was reprecipitated with the addition of 50 μl 13% PEG 8000 and 1.6 M NaCl. After 30 min on ice, the sample was spun at room temperature, and the pellet was washed in 70% ethanol and resuspended in 40 μl TE.

We screened all individuals for polymorphisms in each of three CAPS loci, *Adh*, *Gpa1*, and *Dhs1* (Konieczny and Ausubel 1993), as well as the mitochondrial locus, *Nad5*. Descriptions of the loci are given in Table 2. The nuclear genes were chosen because they have a substantial fraction of noncoding sequence and because they are unlinked; *Adh* is located on chromosome *1*, *Gpa1* is located on chromosome *2*, and *Dhs1* is located on chromosome *4*. Primer sequences are as reported in Konieczny and Ausubel (1993) for the CAPS loci and, for

TABLE 2

Loci surveyed in this study

Gene (enzyme)	Map position	Accession no.	Length	Effective no. of sites	Forward	Position	Reverse	Position
Adh (alcohol dehydrogenase)	1-37	M12196	2084	414	gaaaccaaaa gcattcgatg	601	ggcctttgattaca tgctga	2686
Dhs1 5′ flanking[a,b] (3-deoxy-d-arabino-heptulosonate y-phosphate synthase)	4-8.7	B. Keith, unpublished data	1668	282	caagtgacctga agagtatcg	1	agagagaatgag aaatggagg	1668
Gpa1[a] (G protein α subunit)	2-56.2	M32887	1594	312	gggatttgat gaaggagaac	791	attccttggtctcc atcatc	1398
Nad5 (NADH dehydrogenase subunit 5, exons d and e)	mtDNA	X60048	1916		tcctttcgcgagt cgatacc	993	tcctggcaagctcc tccagt	2910

[a] The amplified regions and primers are identical to those given in Konieczny and Ausubel (1993).
[b] The 3′ base of the amplified fragment corresponds to position 36 in GenBank accession number G166687.

Nad5, the primer sequences are tcctttcgcgagtcgatacc (forward primer) and tcctggcaagctcctccagt (reverse primer). Amplified fragments ranged in size from 1594 (*Gpa1*) to 2084 (*Adh*).

PCR reactions were carried out in 100-μl volumes containing 0.125 mm of each deoxynucleotide, 0.5 μg of each primer, 2.5 units Taq polymerase, 2 mm $MgCl_2$, and 50–100 ng genomic DNA. Conditions for the amplification of *Nad5*, *Adh1*, and *Dhs1* were 3 min at 95°, then 35 cycles of 95° for 30 sec, annealing at 56° for 30 sec, polymerization at 72° for 3 min, followed by extension for 5 min at 72°. The amplification of *Gpa1* used an annealing temperature of 51° but was otherwise identical. The PCR products were phenol:chloroform extracted and cut with each of the eight four-base–recognizing restriction enzymes (*Alu*I, *Dde*I, *Hae*III, *Hin*fI, *Mbo*I, *Mse*I, *Rsa*I, and *Taq*I).

Two methods were used to detect polymorphisms. For the first set of populations in our study (NC6, NC7, RP, Dem, BG, NFC, and NFE), we pooled the four amplicons for each line before restriction enzyme digestion, separated the digested fragments on a 5% denaturing polyacrylamide gel, and transferred the DNA electrophoretically to a nylon membrane (Church and Gilbert 1984). For each locus, a probe was constructed by amplifying the locus from Columbia (Col-0) DNA and purifying the amplicon on an agarose gel. A single strand of the amplicon was uniformly labeled with ^{32}P by incubating 10 ng of denatured, gel-purified, amplified DNA with a single internal primer in a 100-μl reaction mixture containing 10 ng primer, 1.2 units Taq, 5.5 mm, ^{32}P dA, 5 mm dA, 10 mm dGTC, and 1 unit DNA polymerase for 30 min at 37°. Hybridization of each probe to membranes, wash conditions, and autoradiography were as described in Kreitman and Aguadé (1986). Probes were removed from the filters between successive hybridizations by incubating the filters at 65° in a 10 mm Tris-EDTA solution containing 50% formamide. For the remaining populations, we separated our digested PCR products for each locus on a 4% Metaphor agarose gel (FMC, Rockland, ME) in 1× TBE after letting the gel set in the refrigerator overnight. Gels were run at 4.5–6.0 V/cm for ~4 hr and then stained directly with SYBR Green Nucleic Acid Gel Stain (Molecular Probes Inc., Eugene, OR) according to the manufacturer's instructions. All gels (acrylamide and agarose) contained Col-0 as a control.

By comparing across restriction digests, it was possible to distinguish whether RFLP was caused by the loss or gain of a restriction site by a nucleotide substitution, or whether it was caused by an insertion or deletion. Ambiguities were resolved by determining the DNA sequences of the regions in question in the appropriate lines. In addition, we also sequenced through restriction site losses or gains to identify the specific change(s) in lines in which either recombination or parallel substitution was suspected.

Data analysis: To estimate the level of nucleotide variability per site at each locus, we calculated the effective number of sites scrutinized by the four-cutter enzymes in this study (see Kreitman and Aguadé 1984). This number, given in Table 2, represents an estimate of the number of nucleotide sites in the reference sequence (Col-0) that, if mutated, would be detectable as polymorphisms in our experimental system. Nucleotide diversity, π (Nei 1987), was calculated for each locus by taking the average of the observed number of nucleotide differences between all pairs of sequences, either within populations or between populations, and dividing this number by the effective number of sites. Only one allele per individual was used for these calculations, including the one instance where a heterozygote was observed. This estimate of nucleotide diversity, therefore, reflects between-individual variability only.

The selfing rate, s_H, was estimated for the one population

TABLE 3

Polymorphic sites

Locus	Site no.	Mutation[a]	Position	Coding[b] (cd) or Noncoding (ncd)
Adh	1	del (6)	836–927	ncd
	2	*Mbo*I	1303	cd (r)
	3	del (17)	1377	ncd
	4	*Alu*I	1452	cd (r)
	5	*Alu*I	~1761, 1856	cd
	6	ins (2)	1529–1666	?
	7	*Hin*fI	2428	cd (s)
	8	ins (5)	2559	ncd
Dhs1	1	*Rsa*I	88	ncd
	2	*Dde*I, *Hin*fI	167	ncd
	3	*Taq*I	477	ncd
	4	*Alu*I	569	ncd
	5	ins (1) [*Mse*I]	672	ncd
	6	6 del's (13)	672–958	ncd
	7	*Mse*I	~954, 1046	ncd
	8	*Mbo*I	1022	ncd
	9	ins (9) [*Mse*I]	1281	ncd
	10	*Mse*I	1301	ncd
	11	del (4)	1521–1657	ncd
Gpa1	1	*Mse*I	2404–2407	ncd
	2	*Hin*fI	2474–2478	cd
	3	ins (5)	2705–2824	?
	4	del (10)	2705–2948	?
	5	*Mse*I	3403	ncd

[a] Restriction enzyme for which site polymorphism is detected. Insertions (ins) and deletions (del) and their estimated lengths, given in parentheses, relative to Col-0.

[b] Synonymous (s), replacement (r), or unknown (?) change in coding region.

(Kz) in which a single heterozygote was observed at the *Adh* locus. The Kz sample contained 10 homozygous individuals for *Adh* and one heterozygote (alleles 2 and 5 in Table 4). For this calculation, we used the homozygosity estimator of Nordborg and Donnelly (1997)

$$\mathfrak{s}_H = \frac{2(H_w - H_b)}{1 + H_w - 2H_b}, \qquad (1)$$

where H_w and H_b are the proportions of homozygous pairs of alleles (as given in Table 4) within and between individuals, respectively. The allele frequencies given in Table 4 were doubled (except for the Kz heterozygote) to estimate H_w and H_b.

Genealogical relationships of alleles were investigated by analyzing haplotype networks. Haplotype networks were constructed according to the method of Stephens (1985). To construct a network, all haplotypes differing by single changes (a nucleotide substitution or indel) were connected on a graph by a single step. This was repeated for increasing numbers of differences (two, three, etc.) until all haplotypes were connected by their minimum distance. The number of differences between two haplotypes is the number of steps in the shortest path between them in the network. Recombination, parallel mutation, or segregation will create closed loops such that two paths connect particular pairs of individuals, the short path (which represents their true genetic distance) and the

TABLE 4

***Adh* haplotype frequencies**

No.	Haplotype[a]	NC7	NC6	BG	RP	Dem	NFE	NFC	Got	Tamm	Pu2	Kz	Col	La	Tsu	Wu	No	Nd	Ms	TOT
0	10100010	11	10	7	11	10	1	10	10	10	9	4	1	1	1	1	1	1	1	100
1	00100110	0	0	2	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	13
2	10011001	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	7
3	10100011	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
4	11000010	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
5	10011011		0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
Sample size		11	10	10	11	10	12	10	10	10	10	11	1	1	1	1	1	1	1	123

[a] Each variable site is represented by 0 or 1, corresponding to the absence or presence of the site, as described in Table 3. Haplotypes were counted once for each homozygous individual, except for Kz, which contained one heterozygote for alleles 2 and 5.

longer path obtained by connecting each individual to its closest relative. Loops were decomposed into individual variable sites, and for each site in which a parallel mutation was suspected, sequence data were used to further investigate the specific cause.

We examined population differentiation using Holsinger and Mason-Gamer's (1996) hierarchical analysis of nucleotide data. We used their F_{st} estimator

$$\hat{g}_{st} = \frac{\hat{\pi} - \bar{\pi}}{\hat{\pi}}, \quad (2)$$

where $\hat{\pi}$ is total nucleotide diversity and $\bar{\pi}$ is the average within-population nucleotide diversity. g_{st} represents the contribution of variation among individuals from different populations to the total variation. We used the algorithm of Holsinger and Mason-Gamer (1996) to impose hierarchical structure on differentiation among all pairs of populations. Briefly, nodes of a tree were constructed, starting with all populations, by pooling the pair(s) of populations with the smallest pairwise g_{st}. Pairwise g_{st}'s were calculated for the new sample, including the pooled population, and the process was repeated until the entire sample was pooled into a single population. g_{st} at each node of the tree reflects the average relatedness of individuals from the different populations. Critical values for g_{st} were determined from 1000 simulated samples from the pooled populations, and they indicate the probability of observing a g_{st} value as large or larger than that observed if there is no population subdivision.

RESULTS

Types of variability: A relatively small number of polymorphisms were revealed by the four-cutter RFLP analysis. The mitochondrial locus, *Nad5*, exhibited no variability among the 115 field-collected lines and among the seven ecotypes, suggesting a recent common ancestry for the mitochondrial genome or a low mutation rate (Wolfe *et al.* 1987). The remaining analyses, therefore, focus on variability in the three nuclear genes only.

Each of the three nuclear genes contained both nucleotide substitutions and length polymorphisms (indels) in roughly equal proportion (Table 3). Fourteen nucleotide changes were detected in total, four at *Adh*, seven at *Dhs1*, and three at *Gpa1*. Ten indels were detected, four at *Adh*, four at *Dhs1*, and two at *Gpa1*. The ratio of 1.4 nucleotide substitutions to indels in our sample suggests that nucleotide substitutions are approximately sevenfold more common than indels. This conclusion follows from the fact that the four-cutter analysis reveals virtually all indels, but only 19% of nucleotide substitutions for our dataset (see Table 2).

All 10 length variants involved short insertions and deletions, ranging in length from 1 to 17 bp. Indel 6 at *Dhs1* (Table 3), a 13-bp deletion (relative to Col-0), was composed of six small length changes, as revealed by direct sequencing. Only one other indel was larger, a 17-bp deletion at *Adh* (site 3 in Table 3) that was found in the two Eurasian population samples Kz and Pu2.

Variability within individuals: Consistent with *A. thaliana* having a selfing rate close to one, four of the five polymorphic populations in our study yielded only homozygotes. One plant from Kz was heterozygous at a single nucleotide site in *Adh*. This site (site 7 in Table 3) had the most intermediate frequency of any polymorphic site within any of the three loci. This plant, therefore, was probably a result of a recent outcrossing event rather than *de novo* mutation. The selfing rate estimate for the Kz population is still near 1; $\hat{s}_H = 0.91$ for the *Adh* locus, and $\hat{s}_H = 1.0$ for *Gpa1*. The selfing rate could not be calculated for *Dhs1*, which was monomorphic in this population. Since most polymorphisms in the data set are at low frequency, most outcrossing events cannot be observed. Nevertheless, the data are consistent with a very high degree of self-fertilization.

Variability within and among populations: It was possible to unambiguously identify the haplotype of every individual allele at each of the three nuclear loci (Tables 4–6). A total of 24 polymorphic sites (base substitutions and insertion/deletions) in the three loci give rise to only 18 single-locus haplotypes. *Gpa1*, with its three major haplotypes, is the only locus with more than two haplotypes at a frequency >10%. Of the 16 single-locus haplotypes, more than half (10) are represented only once. The majority of the population samples are either invariant at each locus or they have a common haplotype and a single representative of one rare haplotype. Even *Gpa1*, with its three multiply-represented haplotypes, generally had populations fixed for different alleles.

The estimates of within-population nucleotide diversity are consistently low for the three nuclear genes, ranging from 0.00029 for *Gpa1* to 0.00052 for *Adh* (Table 8). With an average intrapopulation nucleotide diversity of only 0.0004, there is very little variability segregating among individuals within local populations.

Nucleotide diversity between populations averaged 0.0014, ranging from 0.0011 to 0.0020 (Table 8). This is approximately four times greater than within-population nucleotide diversity. Estimates of g_{st}, a measure of the betweeen-population component of total variability, range between 0.47 and 0.83, averaging 0.64 for the three loci. Large values of $\hat{g}_{st}$ generally indicate restricted gene flow between populations. In the present case, however, the large estimates of $\hat{g}_{st}$ may be influenced, not only by restricted migration, but also by reduced intrapopulation variability caused by both clonal expansion and selection within local populations.

Genealogical relationship of haplotypes: The low level of within-population nucleotide diversity and the lack of heterozygotes is consistent with the hypothesis that local populations are composed of clonal descendents of a small number of founders. Under such a scenario, genetic associations that can extend between linked loci, and even across chromosomes, will develop. To explore this possibility, we investigated the genealogical relationships of haplotypes to identify potential recombinants within and between loci.

We first constructed haplotype networks for the individual loci (data not shown). *Dhs1* and *Gpa1* haplotypes

TABLE 5

***Dhs1* haplotype frequencies**

No.	Haplotype	NC7	NC6	BG	RP	Dem	NFE	NFC	Got	Tamm	Pu2	Kz	Col	La	Tsu	Wu	No	Nd	Ms	TOT
0	00101100011	11	9	7	11	0	12	10	10	10	9	11	0	1	1	0	0	1	1	104
1	00101101011	0	1	2	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	13
2	00101100010	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
3	10011000101	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
4	01100100011	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
5	00101110011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
6	00011000101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1

TABLE 6

***Gpa1* haplotype frequencies**

No.	Haplotype	NC7	NC6	BG	RP	Dem	NFE	NFC	Got	Tamm	Pu2	Kz	Col	La	Tsu	Wu	No	Nd	Ms	TOT
0	11011	0	7	10	0	10	12	10	0	0	0	9	0	0	0	0	0	1	0	59
1	11110	11	2	0	11	0	0	0	0	10	0	0	0	0	0	0	0	0	0	34
2	11010	0	1	0	0	0	0	0	10	0	8	2	1	1	1	1	1	0	1	27
3	00000	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
4	11000	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1

TABLE 7

Nucleotide diversity estimates

Locus	Nucleotide diversity[a]		
	Within population	Between population	$\hat{g}_{st}$
Adh	0.00052	0.0012	0.63
Dhs1	0.00038	0.0011	0.47
Gpa1	0.00029	0.0020	0.83
Average	0.00040	0.0014	0.64

[a] Average pairwise number of differences / effective number of sites.

produce open networks, indicating that neither recombination nor parallel mutation is required to relate each haplotype one to another. The network for *Adh* indicated the presence of a single recombinant. The putative recombination event was revealed by the presence of four haplotypes based on two site differences, a deletion at position 1377 and an insertion at position 2559. Because these changes involve indels, several different mutations could have led to the band shift. We confirmed by direct sequencing that the haplotypes had identical mutations, suggesting recombination as a plausible mechanism.

The 16 multilocus haplotypes are presented in Table 7. As with the single locus case, the majority of haplotypes (10 of 16) are present only once. More than half (six of 11) of the population samples are fixed for a single multilocus haplotype; these local populations may be composed of clonally related individuals. Of the six multilocus haplotypes that are represented more than once, five are present in two or more populations. Populations belonging to pairs of neighboring sites, NC6–NC7 and NFC–NFE, carry distinctly different haplotypes. Gene flow, while it does not occur at a sufficiently high rate to prevent the differentiation of neighboring populations, has succeeded in distributing the more abundant multilocus haplotypes broadly across the range of the species.

Even though the three loci are located on different chromosomes, the data present few opportunities to identify recombinants. There are five instances, indicated as thin dashed lines in Figure 1, in which a pair of multilocus haplotypes differ by a smaller number of changes than that predicted by the multilocus haplotype network. For instance, the bolded path between the multilocus haplotypes 010 and 011 involves a change from *Dhs1* haplotype 1 to *Dhs1* haplotype 0 at the step designated by the first *b*, followed by the opposite change at the step designated by the second *b*. Because of this reversion, the actual distance between multilocus haplotypes 010 and 011 is two although the cumulative number of changes along the *bolded* path is four. Incongruencies in the network indicate occurrences of parallel mutation, recombination, or chromosomal assort-

TABLE 8

Multilocus haplotype frequencies

No.	Haplotype[a]	NC7	NC6	BG	RP	Dem	NFE	NFC	Got	Tamm	Pu2	Kz	Col	La	Tsu	Wu	No	Nd	Ms	TOT
0	001	11	1	0	11	0	0	0	0	10	0	0	0	0	0	0	0	0	0	33
1	000	0	7	5	0	0	1	10	0	0	0	3	0	0	0	0	0	1	0	27
2	002	0	1	0	0	0	0	0	10	0	7	1	0	1	1	0	0	0	1	22
3	100	0	0	2	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	13
4	010	0	0	2	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	12
5	200	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	6
6	011	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	320	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
8	402	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
9	033	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
10	004	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
11	202	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
12	500	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
13	042	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
14	052	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
15	062	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1

[a] Numbers correspond to a single-locus haplotype, as defined in Tables 4–6. Column 1, *Adh*; column 2, *Dhs1*; column 3, *Gpa1*.

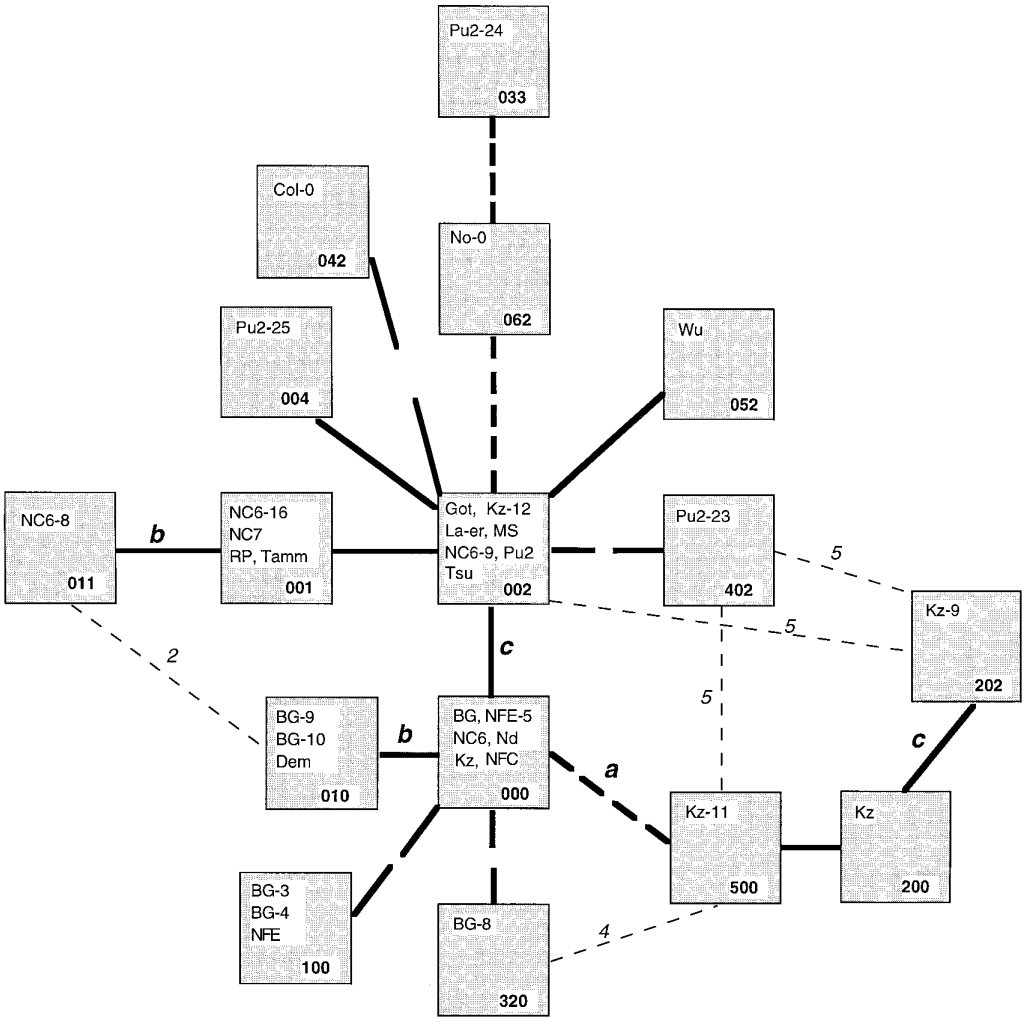

Figure 1.—Multilocus haplotype network for *Adh*, *Dhs1*, and *Gpa1*. Boldfaced numbers at bottom right of boxes are the single-locus haplotypes for *Adh*, *Dhs1*, and *Gpa1*, respectively (see Tables 4–6). Populations having each haplotype are listed in the boxes; for populations containing only rare representatives of a particular multilocus haplotype, the individual line identifications are given. Heavy lines are the number of mutational differences between haplotypes. Thin dashed lines connect pairs of haplotypes differing by a smaller number of changes than that predicted by the number of changes along the network that connect a pair. These incongruencies indicate parallel changes or recombination (chromosome assortment). The numbers above these lines are the number of mutational differences. The italicized letters, *a–c*, denote a minimal set of one intralocus (*a*) and two interlocus (*b* and *c*) recombinational changes that can account for all of the incongruencies in the network (see text for details). Two equally plausible locations in the network are shown for the putative recombination events, *b* and *c*.

ment of haplotypes. By inspection, it is possible to identify a minimum set of three such events that can account for the five incongruencies in the network. One of the events, designated *a* in Figure 1, is the intralocus recombination already identified in *Adh* (sites 6 and 8). The other two events, designated *b* and *c*, involve interlocus recombination (*i.e.*, chromosomal assortment). Event *b* involves recombination between *Dhs1* haplotypes 0 and 1, and *Gpa1* haplotypes 0 and 1. Similarly, event *c* involves putative recombination between *Gpa1* haplotypes 0 and 2, and *Adh* haplotypes 0 and 2. The presence of only two interchromosomal recombination events in the data suggests the possibility of substantial clonal structure within the species and linkage disequilibrium extending across chromosomes. These conclusions, however, are mitigated by the relatively low power of the data to detect linkage disequilibrium.

Relatedness of populations: As previously indicated, polymorphism is approximately four times more common between populations than within populations. In fact, even neighboring populations (NC6–NC7 and NFC–NFE) can be genetically different. From the haplotype analysis, it is apparent that identical haplotypes are widely distributed geographically. To further investigate the genetic relatedness of populations, we used Hol-singer's and Mason-Gamer's (1996) algorithm to impose hierarchical structure on the differentiation patterns. Results for each locus are shown in Figure 2, where the numbers given at each node represent the distance between the two daughter nodes, and the aster-

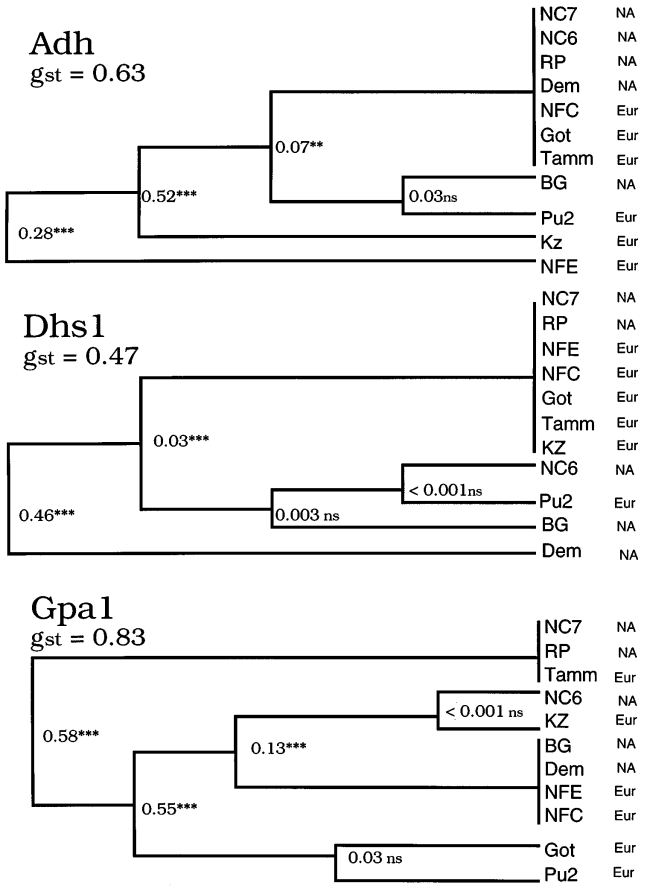

Figure 2.—Hierarchical analysis of haplotype diversity. The number at each node is the distance $\hat{g}_{ij}$ between its two daughter groups. The $\hat{g}_{st}$ value for each locus is the average genetic distance between populations. $^{**}P < 0.01$; $^{***}P < 0.001$.

isks designate significant differentiation between these nodes.

Each of the three loci divide populations into three to four groups that are statistically distinguishable one from another (*Adh* = 3, *Dhs1* = 3, and *Gpa1* = 4). None of the loci reveal any evidence of isolation by distance. On the individual locus trees, North American populations are not more likely to group with North American populations than Eurasian populations, and neighboring populations do not consistently group together. The three loci also indicate rather different relationships among the populations, which suggests an absence of strong linkage disequilibrium between genes on different chromosomes.

DISCUSSION

A. thaliana was found to have low levels of nucleotide variation and substantial population differentiation. The majority of population samples were found to consist of a single prevalent multilocus haplotype. This finding is not unexpected: in the absence of regular long-distance dispersal, a highly self-fertilizing weedy species such as *A. thaliana* is expected to have a patchy distribution of completely inbred colonies. With a selfing rate of ~0.99, any individual heterozygosity will be lost after only a short number of generations. Thus, most individuals are expected to be entirely homozygous. In addition, since regular population extinction and recolonization is expected in an ephemeral species such as Arabidopsis, a small number of inbred founders can contribute to the genetic homogeneity of local populations. Since most seed dispersal occurs over very short distances—likely to be within a meter in Arabidopsis—then these relatively homogeneous patches are expected to remain distinct one from another.

Structure of sampled populations: We have detailed information about many of the populations that we sampled. For example, the Dem (Demotte, IN) population is located in an old, semi-isolated agricultural field several hectares in size. Individual plants could be found across the entire field. The 10 sampled individuals were collected along a transect of ~100 m. This sample consisted of a single multilocus haplotype, so that we cannot reject the hypothesis that these individuals are the direct descendants of a single homozygous ancestor. We do not know, however, whether the whole field consists of this single genotype, or whether the field contains patches of different genotypes. We have ascertained that other Arabidopsis populations in Northern Indiana and Southwest Michigan are polymorphic. Therefore, the homogeneity we found in the Dem sample is not likely to extend geographically much beyond this local population.

We also found different haplotypes in each of two pairs of neighboring population samples. NFC and NFE (Ascot, England) are subpopulations consisting of dense (~2000 individuals per square meter) but disjoint stands of Arabidopsis in a large agricultural field. This field was divided in the past by fences, and the subpopulations have retained the substructure imposed by these fences eight years after their removal (M. Crawley, personal communication). The NFC and NFE subpopulations are ~500 m apart; samples from each were collected from a 10-m by 10-m area. These samples were nearly fixed for different *Adh* alleles, but were identical for *Dhs1* and *Gpa1*. The NC6 and NC7 (Durham, NC) collections were taken from agricultural fields separated by ~1 km. Like the Dem collection, both NC population samples were taken along a transect of 50–100-m length. The NC7 sample contained a single multilocus haplotype, whereas the NC6 sample contained four multilocus haplotypes, the most common of which was different from the NC7 haplotype. Thus, from our data, we can conclude that populations, if defined as single fields, can be variable for multiple genotypes, and neighboring fields and subpopulations can differ substantially both in the alleles present and in their frequencies.

A relatively low interpopulation migration rate is suggested by the presence of distinct haplotypes between

neighboring populations and by our finding that most of the variation is distributed between rather than within populations. Both findings are expected under models of restricted gene flow and extinction/recolonization. Todokoro *et al.* (1996) similarly found that microsatellite variation was distributed between rather than within local Japanese populations of *A. thaliana*. In contrast, individual haplotypes can have worldwide distributions, indicating the importance of long-distance migration in this species.

Comparison of polymorphism levels: Our estimates of interpopulation nucleotide diversity at *Adh*, *Dhs1*, and *Gpa1* are lower than most other published estimates of ecotype nucleotide diversity. The first study providing data on *A. thaliana* polymorphism used 200 λ phage clones as probes in an RFLP analysis of three strains, Niederzenz (Nd-0), Columbia (Col-0), and Landsberg (L*er*; Chang *et al.* 1988). Nucleotide divergence in low-copy number genomic DNA was ~1.4% between Nd-0 and L*er*, 1.3% between Nd-0 and Col-0, and 1.1% between Col-0 and L*er*. While these values are larger than our estimates of nucleotide polymorphism, they overestimate nucleotide diversity for several reasons. First, the estimates are inflated approximately twofold because they are based only on expected (or actual) restriction sites in the data, whereas the estimate of nucleotide polymorphism should also include potential restriction sites (*i.e.*, a site that can form a restriction recognition sequence by a single mutation). The latter constitute approximately half of all nucleotide changes "scrutinized" by restriction enzymes (Hudson 1982). Second, their estimates are based on all band shifts, including insertions and deletions rather than only nucleotide polymorphisms. Third, insertions can contain additional restriction sites, again inflating the number of apparent RFLPs. Fourth, a single length change will be revealed as a polymorphism in every restriction digest, violating the assumption that each band shift represents an independent mutation.

Three other studies estimate nucleotide diversity among ecotypes, all of which yield higher estimates than those presented here and all of which are upward biased. King *et al.* (1993) used 25 λ phage clones as probes to detect polymorphism in 28 ecotypes and found substantial polymorphism. Unfortunately, their calculation of polymorphism was not based on the total number of sites but rather on the number of variable bands only. Thus, it is impossible to quantify the absolute level of variability in their data. In an important study, Konieczny and Ausubel (1993) used RFLPs contained in PCR amplified DNA of the ecotypes Col and L*er* to identify markers that can be used in mapping studies. Amplified DNAs of 18 genes, containing mixtures of intron and coding sequences, were each digested with as many as 83 different restriction enzymes until a polymorphism was revealed. The authors detected 20 polymorphic changes in ~5227 nucleotides, or one base change in 261 bp between the two lines. As in the Chang *et al.* (1988) study, this estimate of nucleotide diversity does not account for potential restriction sites. It also suffers from an ascertainment bias: the authors stopped looking for polymorphism as soon as they found an enzyme that detected one. Therefore, this study also overestimates polymorphism levels.

Another study of sequence polymorphism among ecotypes has recently been carried out for two nuclear CAPS markers, m235 and g2395 (Hardtke *et al.* 1996). Based on a total of 414 bp of comparative data among 18 geographically widely dispersed ecotypes, they identified (approximately) 18 polymorphisms, including base substitutions and indels, with a nucleotide diversity of 0.0221. This estimate includes indels, and therefore overestimates the nucleotide diversity. Reanalysis of their data (GenBank accession numbers Z74001–Z74018), discounting indels, yields a nucleotide diversity estimate of 0.01, a value that is still approximately seven times higher than our estimate of interpopulation nucleotide diversity.

The low levels of variability at the three loci reported in this study are not caused by any systematic bias in four-cutter restriction analysis. First, we are confident that virtually all RFLP variants were scored. Second, in estimating the effective number of sites at each locus, we included only those changes that could be detected on our gels. For example, if a restriction fragment had a potential site near one of its ends that could mutate to form a new restriction site, we would count that site only when the change in the restriction fragment length was sufficiently large to have been detected on our gels. We estimate that 19% of all nucleotide substitutions would have been detected in our study. This leads us to calculate that our study effectively surveyed 1008 bases in the three nuclear loci.

Our estimate of nucleotide diversity for *Adh* may be lower than the true value. Two studies of sequence polymorphism have been carried out for this locus (Hanfstingl *et al.* 1994; Innan *et al.* 1996). Hanfstingl *et al.* (1994) identified a 200-bp hypervariable region in *Adh* between Columbia and Landsberg *erecta*, which differed in this region by 14 nucleotide substitutions (six replacement changes and eight synonymous changes). The remainder of the locus was nearly identical between Col and L*er*. Further analysis of this region among 39 additional ecotypes revealed that all of these ecotypes possessed either the Landsberg-type or the Columbia-type haplotype, with little variation within these haplotype classes. As it turned out, the battery of four-cutter restriction enzymes used in the present study did not detect any of the 14 polymorphisms in the hypervariable region. Thus, at least for this region, nucleotide diversity is underestimated by four cutters. In addition, a more extensive study of *Adh* sequence polymorphism was carried out by Innan *et al.* (1996), who sampled a 2.4-kb region encompassing the locus and 5′ noncoding region

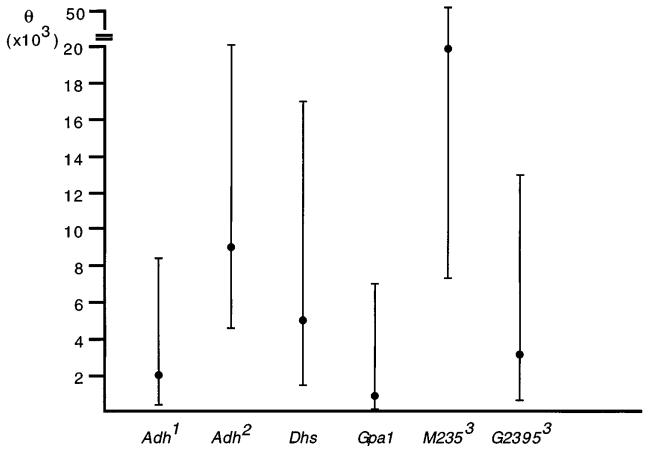

Figure 3.—Estimates of θ for several loci based on the number of segregating sites in each sample. 95% confidence intervals were calculated using Equation 3 in Kreitman and Hudson (1991). The upper and lower limits define values of θ for which there is a 2.5% probability of observing more extreme values. For the three loci reported in this study, a single individual (having the most common haplotype) from each of the 11 population samples and the seven ecotypes were used to estimate θ. [1] from Innan *et al.* (1996); [2] this study; [3] from Hartke *et al.* (1996).

from 17 ecotypes (including L*er* and Col-0). Their estimates of nucleotide diversity are 0.0056 for the coding region (including the hypervariable region) and 0.0080 for the entire 2.4-kb region. Both of these estimates are somewhat higher than our between-population estimate of nucleotide diversity, which we calculate to be 0.0014.

Although our estimates of ecotype nucleotide diversity for *Adh* are lower than those of Hanfstingl *et al.* (1994) and Innan *et al.* (1996), there are also similarities between their studies and ours. In both studies, identical alleles were found to be broadly distributed geographically and there was no evidence of isolation by distance. Both studies found only a small number of haplotypes worldwide. The Innan *et al.* (1996) study found a large proportion of rare mutations (43 of 89 polymorphisms were unique), as well as extensive linkage disequilibrium: only four recombination events can account for six major haplotypes. Innan *et al.* (1996) also found a 5:1 ratio of nucleotide polymorphisms:indels (75:15, respectively) similar to our estimate.

Do any of these estimates of nucleotide polymorphism levels differ significantly from one another? To address this question, we consider a completely neutral model of infinitely many sites with no recombination (Kimura 1969). Under this model, both nucleotide diversity as well as the number of segregating sites in a sample provide unbiased estimators of the neutral parameter $\theta = 4N\mu$. In Figure 3, we have plotted 95% confidence intervals for estimates of θ based on segregating sites. As this Figure shows, all of the 95% confidence intervals overlap, with the single exception of *Gpa1* and M235, the latter having a high level of polymorphism (10 segregating sites in 153 bp). Thus, the estimates of nucleotide diversity based on the loci presented in this study do not systematically differ from those of other loci.

Causes of low levels of polymorphism: There are many reasons to expect low levels of nucleotide diversity in *A. thaliana*. Selfing alone is expected to reduce nucleotide diversity by a factor, $\pi/\pi_0 = (2 - s)/2$, relative to an outcrosser, where s is the selfing rate (Pollack 1987). For *A. thaliana*, where the selfing rate has been estimated to be >0.99, this reduction will be approximately twofold. A more severe reduction in nucleotide diversity is expected to be caused by background selection. The reduction of effective recombination in a highly selfed organism elevates the strength of background selection by enlarging linkage blocks (Charlesworth *et al.* 1993; Nordborg *et al.* 1996). For reasonable estimates of population parameters and deleterious mutation rates, neutral polymorphism can be reduced by at least a factor of 10 in a highly selfed species relative to an outcrosser (Charlesworth *et al.* 1993). Thus, it should not be surprising to find low nucleotide diversity in *A. thaliana*, a highly selfing species.

The influence of population subdivision on nucleotide polymorphism levels is complex. It is clear from this study and others that specific alleles are broadly distributed geographically. Our study suggests that the absolute differentiation between populations is not large, even when comparing populations across continents. This is a strong indication of gene flow (*i.e.*, across continents) or a recent expansion of the species. With polymorphism data for a larger number of independent loci, it will be possible to test an isolation model of population differentiation against the alternative of populations linked by migration (Wakeley 1996). It may be more meaningful, however, to study alternative models for Arabidopsis, such as ones that include local extinction and recolonization with many subpopulations.

A highly selfing organism such as Arabidopsis may be expected to exhibit clonal structure. Measurements of the associations of alleles of loci on different chromosomes allow the possibility of investigating clonal structure. The general lack of variability and the preponderance of a single haplotype at two of the three loci, however, make it difficult to assess whether identical multilocus haplotypes present in different populations are clones or whether they independently arose through segregation. It is certainly the case that our data do not preclude the possibility of widespread clonal haplotypes, and we believe that this deserves further investigation.

Selfing, population subdivision, and background selection are expected to influence nucleotide diversity genome-wide, whereas in the absence of complete clonality, balancing selection and also hitchhiking accompanying selective sweeps will be expected to influence neutral polymorphism levels over smaller genomic in-

tervals. Population structure, like selfing, increases individual homozygosity and therefore decreases effective rates of recombintaion (see Ohta 1982). Subdivision should therefore interact with background selection and selective sweeps to decrease nucleotide diversity within populations. While subdivision is generally thought to increase total nucleotide diversity (Nei and Takahata 1993; Takahata 1994), it can actually have the opposite effect under scenarios of nonconservative gene flow (Nordborg 1997).

In the context of low polymorphism levels genome-wide, a given region with a high polymorphism level is likely to be the result of selection acting to maintain multiple alleles. Both Hanfstingl *et al.* (1994) and Innan *et al.* (1996) suggest the possibility of balancing selection to explain the two distinct *Adh* haplotpes. Similarly, a recent report by Rose *et al.* (1997) of 4.7% nucleotide divergence between Col-0 and L*er* alleles of the *Pat1* locus also raises the possibility of a long-lived balanced polymorphism at this locus. More extensive data sets will be required to determine whether highly diverged alleles are the product of natural selection.

We thank I. Al-Shehbaz, I. Cetl, M. Crawley, R. Mauricio, M. Nachman, G. Robellen, O. Savolainen, and J. Winterer for collecting *A. thaliana* seeds, as well as the Arabidopsis Biological Resource Center at Ohio State University for providing seeds of Arabidopsis ecotypes. B. Keith and H. Ma kindly provided unpublished sequences. Special thanks go to E. Richards for technical assistance. This work was funded by a Packard Fellowship and National Science Foundation Presidential Faculty Fellowship Award DEB-9350363 to J.B. and National Institutes of Health award 1PO1GM50355-01 to M.K.

LITERATURE CITED

Abbott, R. J., and M. F. Gomes, 1989 Population genetic structure and outcrossing rate of *Arabidopsis thaliana*. Heredity **62:** 411–418.

Berry, A. J., J. W. Ajioka and M. Kreitman, 1991 Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. Genetics **129:** 1111–1117.

Chang, C., J. L. Bowman, A. W. DeJohn, E. S. Lander and E. M. Meyerowitz, 1988 Restriction fragment length polymorphism linkage map for *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **85:** 6856–6860.

Charlesworth, B., M. T. Morgan and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289–1303.

Charlesworth, B., M. Nordborg and D. Charlesworth, 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Gen. Res. **70:** 155–174.

Church, G. M., and W. Gilbert, 1984 Genomic sequencing. Proc. Natl. Acad. Sci. USA **81:** 1991–1995.

Cocciolone, S. M., and K. C. Cone, 1993 *Pl-Bh*, an anthocyanin regulatory gene of maize that leads to variegated pigmentation. Genetics **135:** 575–588.

Colosi, J. C., and B. A. Schaal, 1993 Tissue grinding with ball bearings and vortex mixer for DNA extraction. Nucleic Acids Res. **21:** 1051–1052.

Golding, G. B., and C. Strobeck, 1980 Linkage disequilibrium in a finite population that is partially selfing. Genetics **94:** 777–789.

Hamrick, J. L., and M. J. W. Godt, 1989 Allozyme diversity in plant species, pp. 43–63 in *Plant Population Genetics, Breeding, and Genetic Resources*, edited by H. D. Brown, M. T. Clegg, A. L. Kahler and B. S. Weir. Sinauer Associates, Sunderland, MA.

Hanfstingl, U., A. Berry, E. A. Kellogg, J. T. Costa, III, W. Rüdiger *et al.*, 1994 Haplotypic divergence coupled with lack of divergence at the *Arabidopsis thaliana* alcohol dehydrogenase locus: roles for both balancing and directional selection? Genetics **138:** 811–828

Hardtke, C. S., J. Muller and T. Berleth, 1996 Genetic similarity among *Arabidopsis thaliana* ecotypes estimated by DNA sequence comparison. Plant Mol. Biol. **32:** 915–922.

Holsinger, K. E., and R. J. Mason-Gamer, 1996 Hierarchical analysis of nucleotide diversity in geographically structured populations. Genetics **142:** 629–639.

Hudson, R. R., 1982 Estimating genetic variability with restriction endonucleases. Genetics **100:** 711–719.

Hudson, R. R., and N. L. Kaplan 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

Innan, H., F. Tajima, R. Terauchi and N. T. Miyashita, 1996 Intragenic recombination in the *Adh* locus of the wild plant *Arabidopsis thaliana*. Genetics **143:** 1761–1770.

Kimura, M., 1969 The rate of molecular evolution considered from the standpoint of population genetics. Proc. Natl. Acad. Sci. USA **63:** 1181–1188.

King, G., D. Nienhuis and C. Hussey, 1993 Genetic similarity among ecotypes of *Arabidopsis thaliana* estimated by analysis of restriction fragment length polymorphisms. Theor. Appl. Genet. **86:** 1028–1032.

Konieczny, A., and F. M. Ausubel, 1993 A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers. Plant J. **4:** 403–410.

Kreitman, M. E., and M. Aguadé, 1984 Genetic uniformity in two populations of *Drosophila melanogaster* as revealed by filter hybridization of four-nucleotide-recognizing restriction enzyme digests. Proc. Natl. Acad. Sci. USA **83:** 3562–3566.

Kreitman, M. E., and M. Aguadé, 1986 Excess polymorphism at the alcohol dehydrogenase locus in *Drosophila melanogaster*. Genetics **114:** 93–110.

Kreitman, M., and R. R. Hudson, 1991 Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127:** 565–582.

Langley, C. H., J. MacDonald, N. Miyashita and M. Aguadé, 1993 Lack of correlation between interspecific divergence and intraspecific polymorphism at the suppressor of forked region in *Drosohpila melanogaster* and *Drosophila simulans*. Proc. Natl. Acad. Sci. USA **90:** 1800–1803.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23:** 23–35.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nei, M., and N. Takahata, 1993 Effective population size, genetic diversity, and coalescent times in subdivided populations. J. Mol. Evol. **37:** 240–244.

Nordborg, M., 1997 Structured coalescent processes on different time scales. Genetics **146:** 1501–1514.

Nordborg, M., B. Charlesworth and D. Charlesworth, 1996 Increased levels of polymorphism surrounding selectively maintained sites in highly selfing species. Proc. R. Soc. Lond. Ser B **263:** 1033–1039.

Nordborg, M., and P. Donnelly, 1997 The coalescent process with selfing. Genetics **146:** 1185–1195.

Ohta, T., 1982 Linkage disequilibrium due to random genetic drift in finite subdivided populations. Proc. Natl. Acad. Sci. USA **79:** 1940–1944.

Pollack, E., 1987 On the theory of partially inbreeding finite populations. I. Partial selfing. Genetics **117:** 353–360.

Redei, G. P., 1975 *Arabidopsis* as a genetic tool. Annu. Rev. Genet. **9:** 111–127.

Reiter, R. F., R. M. Young and P. A. Scolnik, 1992 Genetic linkage of the *Arabidopsis* linkage map: methods for mapping with recombinant inbreds and random amplified polymorphic DNAs (RAPDs), pp. 170–190 in *Methods in Arabidopsis Research*, edited by C. Konez, N.-H. Chua and J. Schell. World Publishing Co., Singapore.

Rose, A. B., J. Li and R. L. Last, 1997 An allelic series of blue fluorescent trp1 mutants of *Arabidopsis thaliana*. Genetics **145:** 197–205.

Stephens, J. C., 1985 Statistical methods of DNA sequence analysis:

detection of intragenic recombination or gene conversion. Mol. Biol. Evol. **2:** 539–556.

Takahata, N., 1994 Repeated failures that led to the eventual success in human evolution. Mol. Biol. Evol. **11:** 803–805.

Todokoro, S., R. Terauchi and S. Kawano, 1996 Microsatellite polymorphisms in natural populations of *Arabidopsis thaliana* in Japan. Jpn. J. Genet. **70:** 543–554.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. Van De Lee *et al.*, 1995 AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. **23:** 4407–4414.

Wakeley, J., 1996 Distinguishing migration from isolation using the variance of pairwise differences. Theor. Pop. Biol. **49:** 369–386.

Wayne, M., and M. Kreitman, 1996 Reduced variation at *concertina:* a heterochromatic locus. Genet. Res. **68:** 102–108.

Wolfe, K. H., W.-H. Li and P. M. Sharp, 1987 Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. USA. **84:** 9054–9058.

Communicating editor: J. Hey